Using Data Mining to Characterize DNA Mutations by Patient Clinical Features

Steven Evans, MS[a,b], Stephen J. Lemon, MD, MPH[a], Carolyn Deters, RN[b], Ramon M. Fusaro, MD, PhD[a], Carolyn Durham, RN, BSN[a], Carrie Snyder, BSN, OCN[a], and Henry T. Lynch, MD[a]

[a]Hereditary Cancer Institute
Creighton University School of Medicine
Omaha, Nebraska 68178
and
[b]OncorMed, Inc.
Omaha, Nebraska 68102

In most hereditary cancer syndromes, finding a correspondence between various genetic mutations within a gene (genotype) and a patient's clinical cancer history (phenotype) is challenging; to date there are few clinically meaningful correlations between specific DNA intragenic mutations and corresponding cancer types. To define possible genotype and phenotype correlations, we evaluated the application of data mining methodology whereby the clinical cancer histories of gene-mutation-positive patients were used to define valid or "true" patterns for a specific DNA intragenic mutation. The clinical histories of patients with their corresponding detailed attributes without the same oncologic intragenic mutation were labeled incorrect or "false" patterns. The results of data mining technology yielded characterizing rules for the true cases that constituted clinical features which predicted the intragenic mutation. Some of the initial results derived correlations already independently known in the literature, adding to the confidence of using this methodological approach.

INTRODUCTION

At the global level, a significant cancer family history predicts certain gene mutations.[1,2] For example, a strong family history of breast cancer implies possible mutations in the BRCA1 or BRCA2 genes, while a strong family history of colon cancer suggests possible mutations in mismatch repair genes such as MLH1 or MSH2.[3-5] Cancer family history backgrounds have already been widely used by genetic counselors in assisting patients' decisions regarding genetic testing. Beyond the global level, strong research interest exists in discovering patient characteristics which may predict specific intragenic mutations within particular genes and vice versa.

For example, there are dozens of known, clinically pertinent, specific mutations in the BRCA1 gene, and numerous other genes have equally many pertinent intragenic mutations.[6] There is a basis of expectation that intragenic mutations can be characterized by clinical manifestations. For instance, the work of Gayther et al. suggest intragenic truncating mutations within the first two-thirds of the BRCA1 gene (a breast cancer gene) give rise to an excess of ovarian cancer compared to mutations in the last third.[7] This finding remains an open question since research by Serova et al. was unable to confirm the Gayther findings.[8] Overall, there is to date very little replicated research that links specific intragenic mutations within individual cancer susceptibility genes to particular clinical cancer family history presentations.

However, such links would prove quite useful since, for any given individual, testing procedures attempt to find a single mutation among the known mutations in a gene. A single mutation, if found in a given family, is sufficient since there is only the most remote probability that there is more than one significant mutation running in that family. Thus laboratory testing strategies would be far more cost-effective if they could take advantage of clinical data to narrow initially the search among all the known mutations in a gene to just a few mutations with the highest probabilities of concordance. The more these particular intragenic mutations could be accurately predicted and confirmed at very low cost, the more genetically susceptible patients would benefit.

Conversely, knowledge of a specific intragenic mutation correlated to a phenotypical presentation may assist the clinician and genetic counselor in predicting future cancer occurrences in an individual or family. It is also possible that patients' knowledge of the extremely high prospects for a significant intragenic mutation might aid them in making their decision of whether to elect to undergo gene testing.

1091—8280/97/$5.00 © 1997 AMIA, Inc.

In addition, connecting some medical characteristics to mutational data may permit deeper insights into the functional or dysfunctional implications of the various mutations of the genes as well as permit possible identification of other associated genes that may be involved with the disease.

Thus the desired outcome of this research effort was to correlate specific clinical features from patients' cancer family histories with particular DNA intragenic mutations. This basic challenge can be described as characterizing genotypical results (i.e., specific gene mutations) in phenotypical terms (i.e., patient characteristics). Our evaluation of using the methodological approach of data mining to solve this problem is described below.

METHODS

We applied the methodology of data mining to define mutational outcomes in terms of patient attributes. The methodology of data mining is a computational method which analyzes a set of patterns comprised of values (descriptive as well as numerical) assigned to attributes describing the pattern.[9-10] The algorithm of data mining applied in this research in turn is derived from the theory of rough sets, which is a formal mathematical framework for the discovery and representation of regularities in a data set. Essentially rough set methodology provides rigorous mathematical techniques to evaluate the effect of data representation as it pertains to the detection of patterns in the data, particularly imprecise and noisy data.[9-10] At a brush-stroke level, this approach may be distinguished from fuzzy set theory which imposes a numerical metric on imprecise linguistic concepts or human judgments, or neural nets which identify functional relationships in specifically numerical data.

Data mining software analyzes a set of information presented typically in a matrix array, in which each row of the matrix represents an example or instance of the phenomenon of interest and each cell within the row (i.e., a column of the matrix) represents some descriptor of the phenomenon. For example, each row may represent a patient, with the first column indicating which disease (if any) the patient has had, the second column indicating what age he or she had the disease, the third indicating whether the patient is a smoker ("yes" or "no" value), etc. For certain rows (viewed as patients), these rows are marked "true" examples of any phenomenon of interest on the part of the user of the software (e.g., certain patients are identified as true cases of "hereditary breast cancer," or true cases of "patients who survived some disease five or more years," or true cases of patients "who do not buy life insurance," etc.).

Data mining software selects constituent elements (column values) for each row so that one or more boolean logical relationships among these elements characterizes (i.e., predicts) all the "true" examples that were marked, as noted above. One can select which column values the software may utilize to construct possible rules. Through an efficient yet exhaustive process that is complete and systematic, data mining algorithms can find valid rules or patterns that can be constructed so that one or more of the "true" cases are characterized by that rule. The set of all such rules taken together define or characterize the set of "true" patterns provided, and as such, may be thought of as an expert rule-based system which defines or characterizes the patterned "true" set selected.

In our application, the pattern of interest is the cancer family history and associated attributes of a patient population who have undergone gene testing and have tested positive for some specific intragenic mutation in a specific gene. We focused on seven key clinical attributes (summarized in Figure 1) to characterize a cancer family history pattern. Although other researchers might identify a different set of attributes, these seven were distilled from over thirty-five years of clinical experience as well as the cancer genetics literature.[11-12]

The first is the types of cancers the patient and the patient's relatives have had. The values over which this attribute may range is approximately 40 different cancers of anatomic sites and organ systems. The second attribute is the relationship of relatives with cancer to the patient, which essentially is either a first or second degree relative depending upon whether they are a mother, father, or sibling (first degree relatives), or an aunt, uncle, or grandparent (second degree relatives). The third attribute is the age of onset of the cancers, which is a numerical value usually between 1 and 100. The fourth attribute is whether there is evidence of vertical transmission of cancer through generations, with values of 1, 2, or 3 to indicate if one, two, or three or more of the patient's family history generations have had cancer. The fifth attribute is similar to the prior one, except the question is whether there are the same cancers within the same generation, with values 1, 2, and 3 also. The sixth attribute addresses whether the identical type of cancer has occurred within the patient's family history, irrespective of where in the family tree it occurs (with similar values 1, 2, and 3 depending on how many similar cancers are present). The seventh attribute is the extent to which one side of the family or the other (the paternal or maternal side) exhibits a great deal of cancer of any type, with values of 1, 2, and 3 depending upon whether less than one-third of the group has had any type of cancer, more than one-third, or more than two-thirds.

When the analysis process was completed, certain very characteristic patterns of multiple primary cancers arose which can be associated with individual gene mutations (e.g., more frequent occurrences of ovarian and prostate cancer in relatives of a patient with breast cancer). This pattern of highly characteristic, associated cancers in effect comprised an additional characteristic, derived from the initial seven attributes.

Key Attributes of A Cancer Family History

1. Itemized cancers among the patient's relatives
2. Relationship of cancer-affected relatives to patient
3. Age of onset of cancers
4. Evidence of vertical transmission
5. Evidence of cancer in the same generation
6. Repetition of identical cancers in the family
7. Level of overall cancer occurrences

Figure 1 - Selected Features of Hereditary Cancer

One should note that the key clinical attributes we employed cannot be expected to function with full perfection. A number of additional unavailable descriptive factors may be germane such as the ethnic background of the patient. For example, Ashkenazi Jewish women have been found to have an approximately 1% prevalence of the 185delAG mutation in the BRCA1 gene.[13] Still other mitigating circumstances may be present, limiting the value of the information collected, such as reduced penetrance of the germ-line mutation, limited data regarding the actual cancers in the patient's family, the true ages of cancer onset, or confounding factors such as false paternity, etc. Hence any results obtained may be constrained due to the inherent obfuscating factors that inevitably arise in the analysis of cancer family histories. Our research implicitly advances the normative position that the seven attributes presented constitute a practical and productive set which will yield useful characterizations.

As previously noted, in the general application of data mining methodology, selected cases are marked valid (or "true") if they represent the pattern of interest. Otherwise, they are marked as incorrect examples of the pattern (or labeled "false"). The process considers the two sets in such a way as to construct those rules which distinguish the true from the false examples of the pattern of interest. The rules are comprised of one or more selections from the key attributes together with their range of values so that such assignments constitute rules that define or characterize the true set. In our particular application, we had the data from DNA genetic testing on those patients and their family histories for which specific mutations were detected. Thus for a specific positively detected mutation, mutation 5382-insertion C in the BRCA1 gene, we had confidential patient histories in terms of their specific attributes and the particular positive or negative results from their genetic tests for the specific mutation of interest. The cancer family histories of patients with positive mutation test results for a mutation of interest defined the valid or true set. The remaining patients' cancer family histories of individuals who tested positive for some other mutation in the gene but not the mutation of interest were defined to be incorrect or negative examples of patient history attribute data.

RESULTS

We collected the data of family histories of patients who manifested a variety of mutations in the BRCA1 and BRCA2 genes which yield breast and/or ovarian cancers. For our first focus, we defined patients who exhibited a 5382-insertion C mutation in BRCA1 as true (N=40); patients who carried some mutation other than 5382-insertion C were labeled as false (N=505). Using our own data mining technology[14] to derive rules, the following results were obtained:

(#1) all rules contained the attribute "early age of onset" with very high values

(#2) nearly all rules indicated a rather intense family tree of breast cancers, with typically more than 3 cases within one generation of each other.

One of the actual rules takes the form of:

Ca32a=4 **AND** 5< Early < 6 **AND** Gen=2

where *Ca32a* is the code for first degree relatives with breast cancer (which in this rule must have at least four such cases). *Early* refers to early onset of breast cancer, where three points are assigned if any case arises at or before the age of 35, two points are assigned if a case arises between age 36 and age 45, and one point is assigned for a case arising between 46 and age 50. *Gen* refers to the presence of cancers in the same generation, which for this rule requires at least two cases in the same generation. Thus results #1 and #2 above provide specific phenotypical characteristics that predict genotypical results, applicable to the case of patients with a positive mutation 5382-insertion C.

In subsequent runs, we tested (from our data set) four other BRCA1 mutations: 185delAG (N=30), exon 5 missing (N=20), 4808 (N=46), and inferred regulatory mutation (N=12). Rule results were successfully

obtained on these four additional BRCA1 mutations (the two rules for mutation 5382-insertion C noted above remained a unique characterization for that mutation, although this outcome must be re-evaluated when larger values of N become available). The sample size for other specific mutations were smaller than these and were not evaluated; for the same reason, individual BRCA2 mutations were also not yet evaluated. In the general case to date, we obtain a variety of rules, one of which may emphasize very early onset of breast cancer in concert with ovarian cancers in siblings, another may emphasize vertical transmission together with a high repetition of related cancers, etc.

As a second effort, we declared all patients with any BRCA1 mutation as true (N=415), and any patient with any BRCA2 mutation as false (N=130). The rules derived would differentiate BRCA1 from BRCA2 patients. The results of this study produced a single, differentiating result:

If there are one or more first degree ovarian cancers in the cancer family history, then a BRCA1 mutation should be considered first.

DISCUSSION

The data mining approach as we applied it yielded criteria which could depict intragenic mutations based on selected cancer family history attributes. In looking at the data for the result concerning mutation 5382-insertion C which the data mining method utilized, we found that 80.0% of all the cases were characterized by results #1 and #2 summarized above. That is, 80% of the genotype (mutation 5382-insertion C) will be described by either results #1, #2 (or possibly both), and we would predict a BRCA1 5382-insertion C genotype with an 80% confidence level. Although this outcome suggests that the rules might lead to a correct guess 80% of the time, it is tempered by the fact that this is the data utilized by the method itself, and the method may have created a tautology for the data set under review.

Moreover where we have obtained results which could be corroborated with independent results in the published literature, our sample sizes have most often been N=70 or higher. For example, in the case of the latter result differentiating BRCA1 from BRCA2 (N=415), it has already been determined in the literature that ovarian cancer is far more highly correlated with BRCA1 than BRCA2.[15] Thus, although this particular experiment did not yield new scientific information, the fact that we derived already known and independently developed data lends credibility to the approach undertaken. The key recognition is that this approach converges with increasing reliability and confidence to highly significant results as the input database expands.

What is attractive about this approach is that the characteristics predicting specific mutations are cast in both clinical terms as given by a patient's cancer family history as well as by descriptive and/or numerical values assigned to the different characteristics. Thus definitive information is provided which researchers can use to guide further etiologic investigations based on the attributes involved and their relationship to the gene at issue, the cancers under consideration, and other known aspects about the disease.

As we have applied this method to a wide variety of intragenic mutations as well as to an increasingly larger set of mutations for single genes (e.g., all mutations for BRCA1), we have begun to derive a distinction between those clinical attributes which are more significant in their predictive value and those less so. Our expectation is that this could lead to a more normative-free classification of major and minor criteria pertinent in the assessment of hereditary cancer. Such criteria could ultimately be the basis for a classification scheme to aid in the diagnosis and management of hereditary cancers.

In contrast to the method presented, an approach that used neural networks does not provide the explanatory component that data mining methodology yields.[16-17] In addition, a strictly rule-based expert system cannot automatically take advantage of the torrent of data that is forthcoming from ongoing gene testing.[18] In contrast, data mining thrives on the benefit of increased data since this permits more precise refinements between the true and false cases.

Placed in the context of the total cancer burden in the population which surpassed 1.3 million new cases last year, the estimated hereditary component of 5-10% implies 65,000 to 130,000 hereditary cases per year.[19] Given the prospects for testing a significant cohort of these individuals in the future as well as relatives of affected patients, such prospects argue for initiatives as we have presented to achieve cost-effective testing strategies.

CONCLUSION

The emerging area of molecular medicine stands to benefit from added insights into the correlation of those patient characteristics that predict specific genetic mutations as well as specific mutations within various genes. The application of data mining to laboratory test data yields useful results that can build a body of

knowledge about the underlying patient clinical picture, using clear and discernible clinical attributes, which can characterize intragenic mutations.

These data also permit us to predict likely occurrences of specific cancer patterns for disease-unaffected individuals with positive intragenic mutation test results. Such knowledge could be applied by the clinician to significantly aid cancer screening and cancer prevention.

References

1. Lynch HT. Cancer and the family history trail. *NY State Jour of Med* April, 1985; 145-7.
2. Lynch HT, Fitzgibbons RJ Jr, and Lynch JF. Heterogeneity and natural history of breast cancer. *Surg Clin Of NA* 1990; 70:753-74.
3. Wooster R, Neuhausen SL, Mangion J, et al. Localization of a breast cancer susceptibility gene, *BRCA2*, to chromosome 13q12-13. *Science* 1994; 265:2088-90.
4. Fishel R, Lescoe MK, Rao MRS, et al. The human mutator gene homolog MSH2 and its association with hereditary nonpolyposis colon cancer. *Cell* 1993; 75:1027-38.
5. Bronner CE, Baker SM, Morrison PT, et al. Mutations in the DNA mismatch repair gene homologue hMLH1 is associated with hereditary non-polyposis colon cancer. *Nature* 1994; 368: 258-61.
6. Shattuck-Eidens D, McClure M, Simard J, et al. A collaborative survey of 80 mutations in the BRCA1 breast and ovarian cancer susceptibility gene: implications for presymptomatic testing and screening. *JAMA* 1995; 273:535.
7. Gayther SA, Warren W, Mazoyer S, et al. Germline mutations of the BRCA1 gene in breast and ovarian cancer families provide evidence for a genotype-phenotype correlation. *Nature Genet* 1995; 11:428-33.
8. Serova O, Montagna M, Torchard D, et al. A high incidence of BRCA1 mutations in 20 breast-ovarian cancer families. *Am J Hum Genet* 1996; 58:42-51.
9. Pawlak Z. Rough Sets: *Theoretical aspects of reasoning about data*. Dordrecht, The Netherlands: Kluwer Academic Publishers., 1991.
10. Slowinski R. *Intelligent decision support: handbook of applications and advances of the rough sets theory*. Dordrecht, The Netherlands: Kluwer Academic Publishers, 1992.
11. Schneider KA. *Counseling about cancer*. Boston, MA: Dana-Farber Cancer Institute, 1994.
12. Lynch HT, Hirayama T. *Genetic epidemiology of cancer*. Boca Raton: CRC Press, 1989.
13. Struewing JP, Abeliovich D, Peretz T, et al. The carrier frequency of the BRCA1 185delAG mutation is approximately 1 percent in Ashkenazi Jewish individuals. *Nature Genet* 1995; 11:198-200.
14. Evans S, inventor. Methods for identifying human hereditary disease patterns. US patent 5642936, 1997, July 1.
15. Rubin SC, Benjamin I, Behbakht K, et al. Clinical and pathological features of ovarian cancer in women with germ-line mutations of BRCA1. *N Engl J Med* 1996; 335:1413.
16. Wu Y, Giger ML, Doi K, et al. Artificial neural networks in mammography: application to decision making in the diagnosis of breast cancer. Radiology 1993; 187: 81-87.
17. Hassoun MH. *Fundamentals of artificial neural networks*. Cambridge, MA: MIT Press, 1995.
18. Liebowitz J. *Introduction to expert systems*. Santa Cruz, CA: Mitchell Publishing, 1988.
19. Cancer Statistics 1996. *Cancer* 1996; 46: 8.